International Journal of Engineering & Technology, 7 (3.28) (2018) 58-63

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Comparing Generalized Linear Model of Count Data: Application towards Pteridophyta species

Mohd Asrul Affendi Abdullah, Siti Afiqah Muhamad Jamil[1]*, Faridah Kormin[1], Mustafa Mamat[2]

[1]Faculty of Applied Science and Technology, Universiti Tun Hussein Onn Malaysia, Bandar Universiti, 84600, Pagoh, Johor, Malaysia
[2]Faculty of Computing and Applied Mathematics, Universiti Sultan Zainal Abidin, Besut Campus, Terengganu, Malaysia
**Corresponding author E-mail: hw150021@siswa.uthm.edu.my*

Abstract

Pteridophyta is known as “paku-pakis” in Malay and it is one the flora species that exists in ecological system. Besides, *Pteridophyta* is one of the species that need to be preserved. Either flora or fauna, both are very important to preserve the ecosystem and control the pollution. In order to observe the species, the fundamental unit of all diversity metrics is a count of specific individuals. In some consequences, the uncorrected counts of observed species often used in measuring the diversity which ignore detection together and established methods to be used to account for missed species. Analysis of count data is widely used in engineering, public health, epidemiology, medical studies, ecology and many research of interest. Rarity increases the number of locations with zero detection in excess of those expected under simple models of abundance. The aims of this study are to compare the Generalized Linear Model (GLiM) in the application of *Pteridophyta* species of count data.

Keywords: *Poisson model; Zero-Inflated Poisson; Negative Binomial; Akaike Information Criterion (AIC).*

1. Introduction

Conservation of floral is important of floral is important in stabilizing the biodiversity of the earth. Consequently, ecology comprises the relation of the organisms to their environment as the plants are useful in reviewing the pattern of biotic response to their climate change and they are abundant and sensitive to the climate variables which are the temperature, seasonality and rainfall [4]. Besides, diversity estimates are central to the community and macro ecology and are frequently used in conversation ecology. The fundamental unit all of the diversity metrics is a count of species, individuals or both. Attempt to define the imperfect detection, the predictable consequences occur when the species are rare, missed individual’s results in false absences. Hence, uncorrected counts of observed species often used in the measure of diversity ignore detection together and established methods used to account for missed species do not disentangle detection from occurrence. Ecological data are normally described as the count observation with their non-negative values. In recent research, assessed the implication count data, the analysis of count data is widely used in medical studies, epidemiology, ecology and many research of interest. The rarity increases the number of locations with zero detection in excess of those expected under a simple model of abundances such as the Poisson regression analysis or the negative binomial regression analysis. Subsequently, in ecology, normally the count data appears as overdispersed and a common approach to deal with it is by the generalized linear model framework [3]. Besides, environmental factors or habitat condition are favorable to their species of interest.

Besides, policy statements of Malaysia’s National Policy on Biological Diversity said: “To conserve Malaysia’s biological diversity and to ensure that its components are utilized in a sustainable manner for the continued progress and socio-economic development of the nation”. Through that statement, as Malaysia covered by tropical floral, the humid tropics attempted to provide a climate that is suitable to support the rich and diverse life forms. Hence, it could sustain the population of flora and fauna on the natural ecological habitats.

The studies of floral kingdom include the non-seed bearing plants which indicate the lower plants and also the seed bearing plants which comprised the higher plants. Focused on the non-seed bearing plants, it covered on the algae, mosses and fern. Instead of that, the seed bearing plants consists of the angiosperms (flowering plants) and the gymnosperm (cone-bearing plants). Explored on the non-seed bearing plants, botanically, the fern is classified as Pteridophyta and they are common as it is edible to be eaten raw, cooked, preserved in brine or pickled. The ferns do not produce through seed but they produced by the spores. Besides, Pteridophyta has true roots and leaves with a short stem and sometimes called as rhizomes. Based on the previous study by [1], the classification of living ferns comprised four classes of order which is 11 orders and 37 families with living representatives of 11,500 species worldwide. Consequently, there are 1,165 species of pteridophytes in Malaysia; 647 species occur in Peninsular Malaysia and 750 in Sabah while 615 in Sarawak [2].

The generalized linear model has been used in analyzing the count observation as several standard statistical analyses of parametric models for non-normally distributed data covers the Poisson regression model, negative binomial regression model, zero-inflated models, and hurdle models [3]. These models have the power of parametric model and more flexible in handling repeated measures, multiple covariates and various structures of fixed and random effects as it has been assumed difference than the normal distribution. Poisson distribution defines as the number of events that occur in a fixed period of time and as the mean count increases, the distribution could be approximately normal. According to [8], to assess the relationship between the abundance of a species in environmental characteristics, the study could focus on the regression methods within the generalized linear model framework. In spite of that, suitable methods could be used in estimating Pteridophyta species under simple model abundance.

2. Methodology

This paper gives an account of Poisson regression analysis, negative binomial regression analysis, zero inflated models for both zero inflated Poisson regression and zero inflated negative binomial regression analysis if necessary. Noteworthy, the model adequacy checking plays an important role in selecting the suitable models which could be used in the study.
Before that, Pteridophyta has been collected very well at certain places in Malaysia for example in Mount Kinabalu. But, in another area, the ferns or the Pteridophyta were not well presented and need to be conserved. In this case, the abundance data of Pteridophyta species could be estimated by this generalized linear model as it represents the appropriate method in analyzing the count outcomes. Thus, further estimation was well presented in this study.

2.1. Poisson regression analysis

In order to perform Poisson regression analysis, this regression model was typically described in terms of their systematic component in which the response was linked to the environmental data. Based on the previous researcher by [8], the study used model structure and model specification as the terms in describing the Poisson model. Explored the model structure, it has the choice of environmental characteristics which were the explanatory variables that assumed to affect the species abundance which was the response variables and the shape of modeled responses such as linear or quadratic model. Therefore, the model specification describes ways in relating these variables by using the 'link' function. Since Poisson used 'link' function, the response variable showed non-linearly related to the explanatory variables. According to [3], from the general function, the variables were linked by the log transformation:

$$\text{Log}\,(p) = \ln(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n$$

where the response variable, p was the probability of an event occurring while X_0 were the independent variables and βn were the regression coefficient by the simple standard Poisson regression model. Besides, a standard framework for describing the Poisson model was the equidispersion characteristics. It means the assumption of mean and variances were assumed to be equal [7] and hence, the distributions were likely to be normal. In the case of violating in Poisson, due to the larger frequency of extreme observations, the results would likely appear as overdispersion with the variance greater than its mean value. In Poisson distribution family, the mean and variances take all possible values from positive infinity to negative infinity which could be described as:

$$\text{F}(y|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma}} \exp - \frac{(y-\mu)^2}{2\sigma^2}$$

Given that $y! = y(y-1)(y-2)\ldots(2)(1)$ and μ was the arithmetic mean number of incidence at specific points of time. The probability of counts depends on the variance:

$$\text{F}(Y = y!\ \mu) = \frac{\mu^2}{y!} \exp(-\mu), y_i = 0, 1, 2, \ldots, \infty$$

Furthermore, Poisson distribution only specifies by one parameter, μ as both mean and variances was equal. In this case of study, the Pteridophyta species were analyzed by using Statistical Package for Social Science (SPSS) version 20 and Microsoft Excel 2010. The GLiM was applied as the dependent variable involved count data of discrete variables. In order to make the study more convincing, the relevant statistical analysis procedure has been included, so that the hypotheses could be achieved.
The logarithm of the response variable was linked to a linear function of explanatory variable such that:

$$\ln(\mu i) = \beta' xi \text{ or } \mu_i = \exp(\beta' xi)$$

On the other hand, a standard Poisson regression model expressed the log outcome rate as a linear function of a set predictors.

2.2. Overdispersion checking

To further illuminate the Poisson regression analysis, the most important part was detecting the presence of overdispersion. If the equality of mean and variances was violated, an overdispersion feature of the problem could occur. There were three conditions which could describe the key of overdispersed model. The first condition indicates, if deviance/df > 1, then, overdispersion might be present. The second condition was, if deviance/df < 1, then underdispersion might occur. Lastly, $\mathcal{X}^2$ with degrees of freedom equal to g (has a mean of g). Benefit for using the maximum likelihood method, the simple likelihood ratio test would be used to assess the adequacy of the negative binomial over the Poisson regression.
In order to overwhelm the problem of overdispersion, the generalized of Poisson which was the negative binomial regression model has been suggested by [10]. Therefore, the value of deviance which was used to calculate the overdispersion could be described as:

$$Deviance\ (D) = 2\left\{\Sigma[y \ln\left(\frac{y}{\hat{\mu}}\right) - (y - \hat{\mu})]\right\}$$

Besides, in ordinary least square regression analysis, similarly, the coefficient of determination, R^2, has been defining as the deviance in analyzing count data. This R^2 value was used to provide the descriptive information about the model fit:

$$R^2 = \frac{\Sigma(\hat{y} - \bar{y})^2}{\Sigma(y - \bar{y})^2}$$

where the observed value was y, the predicted value was $\hat{y}$ with the mean value of y. In addition, the model adequacy checking compares a fitted model to a saturated model. The difference between log-likelihood under two model:

$$D = 2 \ln \lambda = 2[l(b\ max; y) - l(b; y)]$$

where $l(b\ max; y)$ the log likelihood function for the maximum model and $l(b;\ y)$ is the log likelihood function for testing the model. Thus, the hypothesis of the test is:

H_0: The model is a good fit
H_1: The model is not a good fit

The deviance based on $\mathcal{X}^2$ with n-p degree of freedom. The null hypothesis is rejected if deviance, D > $\mathcal{X}^2_{\infty}$ n-p and vice versa.

2.3. Negative Binomial regression analysis

Previously mentioned, if overdispersion exists, there is a possibility for the analysis to use negative binomial regression so that, the results could fit the data very well. Based on [6], if overdispersion exists, there are two properties normally associated with the presence of overdispersion which is about the contagion and through excess zero. “Contagion” could be defined when the individuals are more grouped than expected as they happened individual,y while the excess zero is when the total observations happened to have more zero than what they would expect. In spite of that, both contagion and excess zeros may increase the variance relative to the mean which then contributes to the overdispersion. As overdispersion increased, in the statistical analysis of ecology, the result produce would be more accurate [5, 9]. In addition, negative binomial has been known as the generalized of Poisson regression as the assumption of Poisson which is the variance is not equal to the mean. The calculated value of negative binomial is:

$$\mu_i = \exp(X_i\beta_i + e_i) = \exp(X_i\beta_i)\exp(e_i)$$

where $\exp(e_i)$ ~ Gamma $(\alpha^{-1}, \alpha^{-1})$ and the function of the density can be derived as:

$$f(Y_i|X_i) = \frac{\Gamma(y_i + \alpha^{-1})}{\Gamma(y_i + 1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_i}\right) + \left(\frac{\mu_i}{\alpha^{-1} + \mu_i}\right)y_i$$

where Γ represent the gamma integral which specializes to a factorial integer argument. Besides, the estimation of the parameters is done by maximizing the log likelihood function;

$$\ln L(\beta) = \Sigma_{i=1}^{n}(y_i x_i'\beta - \exp(x_i'\beta - \ln y_i\,!)$$

Therefore, the Maximum Likelihood Estimator is done by the numerical method generated using the computer-based iteration procedure in GenStat [10].

2.4. AIC of regression model

The smallest AIC values can be used to identify the appropriate or suitable method in modeling the observations of counts data. AIC was considered as GLiM models was part of nested models. Otherwise, the analysis must use Vuong test to compare the non-nested models such as the Poisson Hurdle versus negative binomial regression model.

2.5. Incidence rate ratio (IRR) value

Incidence rate ratio was another way in explaining the negative binomial regression model. From the final model which was the negative binomial, IRR indicates some explanation in statistical analysis which displays the best-selected variables and could accurately estimate the numbers of Pteridophyta grows that includes:

1. IRR = 1 (numerator group and denominator group have same incidence rate)
2. IRR > 1 (numerator group has higher incidence rate than denominator group)
3. IRR < 1 (numerator group has a lower incidence rate than denominator group)

3. Results and Analysis

By using SAS statistical software, this study performed the Poisson regression analysis, followed by zero-inflated Poisson regression analysis and negative binomial regression analysis. These models are part of GLiM, which is used to analyse the count data as observation may indicate high zero values and having a non-normal observation. In order to analyse the count data, this study used PROC GENMOD, PROC UNIVARIATE and PROC IMPORT.

3.1. Poisson Regression model

Based on Table 1, the results showed the summarization of the estimation score obtained by the Poisson regression model. Starting from the 95% of confidence limits, standard error and the *P*-value, the variable of Types (types of Pteridophyta species) and the variable of years (duration of Pteridophyta species grows) have been summarized as follow.

Table 1: Parameter estimate of Poisson regression model

Analysis Maximum Likelihood Parameter Estimates			
Parameter	Estimate (95% CI)	Std. Error	P-Value
Intercept	4.8727 (4.7681, 4.9974)	0.0534	<0.0001
Types: ALF	-0.0202 (-0.1595, 0.1191)	0.0711	>0.05
Types: DDF	1.2669 (1.1560, 1.3779)	0.0566	<0.0001
Types: GPF	0.6806 (0.0614, 0.5603)	0.0614	<0.0001
Types: NBF	1.6606 (1.5537, 1.7675)	0.0545	<0.0001
Types: SPF	-0.0780 (-0.2193, 0.0634)	0.0721	>0.05
Types: VEF	0.0000 (0.0707, -0.1386)	0.0707	1.0000
Types: VLF	0.0000 (0.0000, 0.0000)	0.0000	-
Years: 4	-2.1894 (-2.3304, -2.0485)	0.0719	<0.0001
Years: 8	-0.5082 (-0.5812, -0.4352)	0.0372	<0.0001
Years: 12	0.2982 (0.2931, 0.3572)	0.0301	<0.0001
Years: 20	0.0000 (0.0000, 0.0000)	0.0000	-

Based on Table 1, when the probability values indicate less than α = 0.05, the study derived that the variable significantly affected the total numbers of Pteridophyta in that time. For the *P*-value of types of Pteridophyta, DDF, GPF and NDF were < 0.0001 while 4 years, 8 years and 12 years showed *P*-value < 0.0001. Hence, the variables significantly affected the total numbers of Pteridophyta in that area. However, the ALF, SPF, VEF have significantly not affected the species to grow.

$\widehat{\text{Log}\,(y)}$ = 4.8727 -0.0202**X1** + 1.2669**X2+**0.6806 **X3+**1.6606 **X4-**0.0780**X5-**2.1894 **X8-**0.5082X9+0.2982 **X10**

Therefore, the equation of Poisson regression model could be summarized as below.

$\widehat{\text{Log}\,(y)}$ = 4.8727 + 1.2669**DDF+**0.6806 **GPF+**1.6606 **NBF-**2.1894 **years4-**0.5082**years8+**0.2982**years12**

The SAS procedure could be used as the followed.

```
PROC GENMOD DATA=FIQ.PTERIDOPHYTA;
CLASS TYPES YEARS/ PARAM=GLM;
MODEL NUMBERS = TYPES YEARS/TYPE3
DIST=POISSON;
RUN;
```

3.2. Goodness of fit Poisson regression model

By referring to the Table 2, since the *P*-value was 0.000 less that than α with the value of 0.05, the study rejects null hypothesis indicating that the Poisson regression model does not fit the data reasonably well.

Table 2: Goodness of fit by Poisson model

Criterion	Deviance
Df	18
Chisq	948.3119
P-value	0.000

The SAS procedure could be used as the followed.

```
DATA FIQ.PTERIDOPHYTA;
DF=18; CHISQ=948.3119;
PVALUE=1-PROBCHI (CHISQ,DF);
RUN;
PROC PRINT DATA=FIQ.PTERIDOPHYTA;
RUN;
```

Since *P*-value < 0.05, the model does not fit the data. This study checks on the presence of overdispersion and excess zero before proceeding with the other models of GLiM.

3.3. Checking on the overdispersion

Table 3: Overdispersion by deviance model

Criterion	Deviance
Value	948.3119
Value/Df	52.6840

Based on Table 3, overdispersed by deviance, the value of deviance was 948.3119 and the average value of deviance was 52.6840. Therefore, as the *P*-value equals to 52.6840 previously, which was greater than 1. This indicates that overdispersion exists. Checking the overdispersion was compulsory as it was one of the model adequacies checking for the GLiM. Besides, another method in checking the overdispersion was by comparing the mean and value of variance. If the value of variance was greater than the mean, thus, overdispersion exists.

3.4. Checking excess zero observation

By using the following SAS procedure;

```
PROC UNIVARIATE DATA = FIQ.PTERIDOPHYTA
NOPRINT;
HISTOGRAM NUMBERS / MIDPOINTS = 0 TO 50 BY 10
VSCALE = COUNT;
RUN;
```

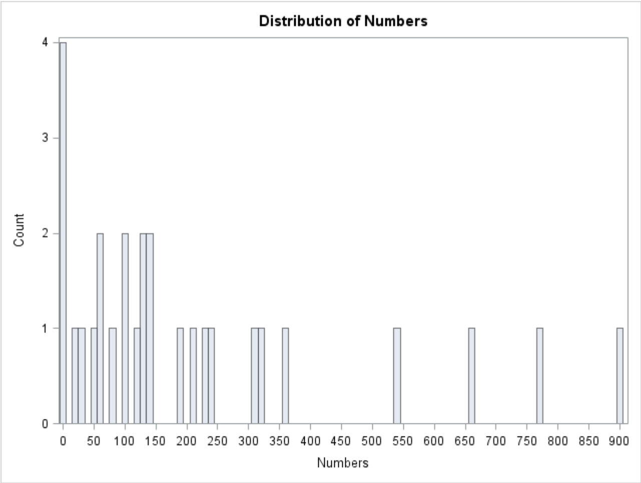

Fig. 1: Excess zero values

Fig. 1 shows that zero value exists with four times of the observations. Since the zero values are not too many, this study needs to perform both zero-inflated Poisson and negative binomial to get accurate results for *Pteridophyta* species.

3.5. Zero-Inflated Poisson regression model

In order to perform zero-inflated Poisson regression model, this study uses the following SAS codes.

```
PROC GENMOD DATA=FIQ.PTERIDOPHYTA;
CLASS TYPES YEARS;
MODEL NUMBERS = TYPES YEARS/ DIST=ZIP;
ZEROMODEL NUMBERS /LINK=LOGIT;
RUN;
```

Table 4 shows the summarization of the estimation score obtained by the zero-inflated Poisson regression model.

Table 4: Parameter estimate of zero-inflated Poisson regression model

Analysis Maximum Likelihood Parameter Estimates			
Parameter	Estimate (95% CI)	Std. Error	P-Value
Intercept	4.9100 (4.8055,5.0145)	0.0533	<0.0001
Types: ALF	-0.0202 (-0.1595,0.1191)	0.0711	>0.0500
Types: DDF	1.2158 (1.1047,1.3270)	0.0567	<0.0001
Types: GPF	0.6295 (0.5090,0.7499)	0.0615	<0.0001
Types: NBF	1.6095 (1.5024,1.7166)	0.0547	<0.0001
Types: SPF	-0.0779 (-0.2193, 0.0635)	0.0721	>0.2000
Types: VEF	0.0000 (-0.1386, 0.1386)	0.0707	>0.5000
Types: VLF	0.0000 (0.0000,0.0000)	0.0000	-
Years: 4	-1.8667(-2.0085,-1.7248)	0.0724	<0.0001
Years: 8	-0.5082(-0.5812,-0.4352)	0.0372	<0.0001
Years: 12	0.2982(0.2391,0.3572)	0.0301	<0.0001
Years: 20	0.0000 (0.0000, 0.0000)	0.0000	-

Based on the following Table 5, variables DDF, GPF, NBF, 4 years, 8 years and 12 years are significant towards the observation of *Pteridophyta*.

Before that, by performing the goodness of fit, this study found that zero-inflated Poisson does not fit the data of *Pteridophyta*.

```
DATA FIQ.PTERIDOPHYTA;
DF=16; CHISQ=572.5962;
PVALUE=1-PROBCHI(CHISQ,DF);
RUN;
PROC PRINT DATA=FIQ.PTERIDOPHYTA;
RUN;
```

Table 5: Goodness of fit by zero-inflated Poisson model

Criterion	Deviance
Df	16
Chisq	572.5962
P-value	0.000

Since *P*-value < 0.05, the model does not fit the data. Then, this study proceeds with negative binomial GLiM.

3.6. Negative Binomial regression model

The SAS code below performs the negative binomial regression analysis.

```
PROC GENMOD DATA=FIQ.PTERIDOPHYTA;
CLASS TYPES YEARS/ PARAM=GLM;
MODEL NUMBERS = TYPES YEARS/TYPE3 DIST=NB;
RUN;
```

Based on Table 6, the results showed the summarization of the estimation score obtained by the negative binomial regression model.

Table 6: Parameter estimate of Negative Binomial regression model

Analysis Maximum Likelihood Parameter Estimates			
Parameter	Estimate (95% CI)	Std. Error	*P*-Value
Intercept	4.7976 (4.0285, 5.5668)	0.3924	<0.0001
Types: ALF	0.1207 (-0.7845, 1.0260)	0.4691	0.7938
Types: DDF	1.8649 (09204, 2.8094)	0.4819	0.0001
Types: GPF	1.0861 (0.1597, 2.0124)	0.4726	0.0216
Types: NBF	2.6087 (1.6442, 3.5732)	0.4921	<0.0001
Types: SPF	0.1545 (-0.7751, 1.0841)	0.4743	0.7446
Types: VEF	0.0670 (-0.8449, 0.9788)	0.4652	0.8856
Types: VLF	0.0000 (0.0000, 0.0000)	0.0000	-
Years: 4	-3.3066 (-4.0972, -2.5159)	0.4034	<0.0001
Years: 8	-0.9391 (-1.6179, -0.2603)	0.3463	0.0067
Years: 12	0.0235 (-0.6628, 0.7098)	0.3502	0.9465
Years: 20	0.0000 (0.0000, 0.0000)	0.0000	-

When the probability values indicate less than α = 0.05, the study derived that the variable significantly affected the total numbers of Pteridophyta in that time. For the *P*-value of types of Pteridophyta, DDF was 0.0001 less than 0.05 and NBF were less than 0.0001. While ALF equal to 0.7938, GPF equal to 0.0216, SPF was 0.7446 and VEF was 0.8856 were all greater than α, 0.05. Besides, 4 years has shown *P*-value less than 0.0001. The 8 years equal to 0.0067 less than 0.05, while for 12 years 0.9465 greater than 0.05.

Hence, those probability values were significant in predicting the total numbers of Pteridophyta if the variables were less than alpha value, 0.05. Overall, variables that were not significant comprised the ALF, SPF, VEF and 20 years. Besides, by checking the goodness of fit of negative binomial model.

```
DATA FIQ.PTERIDOPHYTA;
DF=18; CHISQ=25.1724;
PVALUE=1-PROBCHI(CHISQ,DF);
RUN;
PROC PRINT DATA=FIQ.PTERIDOPHYTA;
RUN;
```

Table 7: Goodness of fit by negative binomial model

Criterion	Deviance
Df	18
Chisq	25.1724
P-value	0.12024

Based on Table 7, since *P*-value = 0.12024 > 0.05, the model does fit the data. Then, this study proved that *Pteridophyta* data fit by using the negative binomial GLiM.

The formulated equation assessed by negative binomial regression model has been shown as below:

Numbers = 121.2191 exp (1.8649***DDF*** + 1.0861***GPF*** + 2.6087***NBF*** - 0.9391***years4*** + 0.0235***years12***)

where exp (***e***) ~ Gamma (α^{-1} ,α^{-1})

3.7. The Akaike Information Criterion (AIC)

Table 8: AIC and BIC scores of both models

Models	Poisson Model	Zero-Inflated Poisson Model	Negative Binomial Model
AIC	1133.9767	1010.2405	327.7790

Based on the results goodness of fit previously by the negative binomial regression model and by observing the Akaike Information Criterion (AIC), the results obviously showed that negative binomial is the most suitable model that fit the *Pteridophyta* data. Based on the smallest values of the GLiM models, a negative binomial regression model fits with the smallest value of AIC and *P*-value greater than 0.05 (95% confidence interval).

3.8. Interpreting the incidence ratio rate (IRR) of Negative Binomial regression

Table 9: Estimated IRR value based on negative binomial model

Parameter	Estimate	Standard Error	IRR
Intercept	4.7976	0.3924	121.223
Types ALF	0.1207	0.4691	1.128
Types DDF	1.8649	0.4819	6.455
Types GPF	1.0861	0.4726	2.963
Types NBF	2.6087	0.4921	13.581
Types SPF	0.1545	0.4743	1.167
Types VEF	0.06695	0.4652	1.069
Types VLF	0	0.0000	1.000
Years 4	-3.3066	0.4034	0.037
Years 8	-0.9391	0.3463	0.391
Years 12	0.02349	0.3502	1.024
Years 20	0	0.0000	1.000

Based on Table 9, starting from ALF, the IRR = 1.128 (12.8%), the total number of Pteridophyta is expected to increase by 12.8% when involving ALF compared to VLF. Then, IRR = 6.455 (545.5%) for DDF, the total number of Pteridophyta is expected to increase by 545.5%% when involving DDF compared to VLF. For GPF, IRR = 2.963 (196.3%) indicating that the total number of Pteridophyta is expected to increase by 196.3% when involving GPF compared to VLF. Besides, for NBF, the IRR = 13.581 (1258.1%) means that total number of Pteridophyta is expected to increase by 1258.1% when involving NBF compared to VLF. Besides, for SPF and VEF, the IRR values were 1.167 (16.7%) and 1.069 (6.9%) respectively. Thus, the total number of Pteridophyta is expected to be increased by 16.7% and 6.9% when involving SPF and VEF respectively compared to VLF. On the other hand, for the duration of plants grows, within 4 years, the IRR = 0.037 (-96.3%) which means that the total number of Pteridophyta is expected to be decreased by 96.3% when involving 4 years compared to 20 years. The IRR for 8 years was equal to 0.391 (-60.9%). That means the total number of Pteridophyta is expected to be decreased by 60.9% when involving 8 years compared to 20 years. For 12 years, the IRR = 1.024 (2.4%) indicates that total number of Pteridophyta is expected to be increased by 2.4% when involving 12 years compared to 20 years.

Hence, the results showed that all types of Pteridophyta seem to have influence in predicting the total numbers of Pteridophyta grows as all types showed increased values of IRR. While, the duration that is suitable for the Pteridophyta were within 12 years and 20 years.

4. Discussion

By approaching the objective of study consequently, the Poisson regression model and zero-inflated Poisson regression model does not fit the data really well compared to negative binomial regression analysis. Before that, by checking the model adequacy, the data showed the presence of overdispersion and this may lead the data to have not normally distributed due to the large extreme value of observations. In spite of that, the study proposed negative binomial regression analysis instead of Poisson regression model and zero-inflated Poisson. By proceeding the negative binomial, the study used AIC value and check the goodness of fit of the model. The outcome indicated that smaller value of AIC was in the negative binomial regression analysis and the model fit the data. Hence, a negative binomial regression model was proposed in analyzing the count data of the total number of Pteridophyta.

Finally, the study used incidence rate ratio (IRR) of the negative binomial regression model in order to interpret the model and at the same time detecting the variable that eventually affects the total numbers of Pteridophyta to grow more efficiently. As a result, all types of Pteridophyta species seems to have influenced in predicting that species. While, 12 years and 20 years were a more suitable time in indicating the high number of Pteridophyta species.

5. Conclusion

The GLiM consists of several statistical analyses of parametric models for non-normally distributed data which includes the Poisson, negative binomial, zero-inflated and hurdle models. In this study, Poisson regression analysis, zero-inflated Poisson and negative binomial were used in predicting the total number of *Pteridophyta* species. Since the data was counted, it was difficult in accurately estimating the values.

Besides, the data was limitedly taken on fixed duration of time which was within 4 years, 8 years, 12 years and 20 years. As the data was secondary data, it was difficult to estimate the certain time needed. In addition, in terms of the counting the total number of plants, this secondary data did not clearly explain the estimated values within what area, the width or how far the observation has been taken. Then, the data was only comprised the categorical data which was within the types of Pteridophyta and the duration of years the plants grow. The study recommended to use other types of the method in predicting the total number of Pteridophyta species and adding more suitable factors to increase the accuracy in predicting the Pteridophyta species.

Acknowledgement

We would like to thank the Ministry of Higher Education, Universiti Tun Hussien Onn Malaysia, for supporting this research project.

References

[1] J. H. Beaman and P. J. Edward, “Fern of Kinabalu: An Introduction Natural History Publication,” 2007, pp. 198.

[2] Y. C. Wee, “Ferns of the tropics, Marshall Cavendish,” 2005, pp. 199.

[3] J. A. Nelder and P. McCullagh,”Generalized Linear Model,” *The Annals of Statistics*, 1983, 12(4), 1589-1596.

[4] M. A. Barton, “Floral Diversity and Climate Change in Central Colorado during Eocene and Oligocene,” *ProQuest LLC,* 2006, pp. 1-68.

[5] J. M. Ven Hoef and P. L. Boyang, “Quasi-Poisson vs. Negative Binomial regression: How Should we Model Overdispersed Count data?” *Ecological Society of America,* 88(11), 2766-2772.

[6] A. N. H. Smith, M. J. Anderson, and R. B. Millar, “Incorporating the intraspecific occupancy-abundance relationship into zero-inflated models,” *Ecological Society of America,* 93(12), 2526-2523.

[7] S. A. Ahmed, J. S. Saddiqqa, Quaiser Saghir and S. Kamal, “Using PCA, Poisson and Negative Binomial Model to study the Climatic Factor and Dengue Fever Outbreak in Lahore,” *Journal of Basic and Applied Sciences,* 11, 6679.

[8] J. M. Potts and J. Elith, “Comparing species abundance models,” *Ecological Modelling,* 2006, 199, 153-16.

[9] S. J. Wenger and M. C. Freeman, “Estimating species occurrences, abundance, and detection probability using zero inflated distributions,” *Ecology,* 89, 2953-2959.

[10] N. E. Breslow, “Extra-Poisson Variation in Log-Linear Models,” *Journal Royal Statistical Society,* 1984, 33(1), 38-44.